語形と文字種による未知語の品詞の推論

1Q－7

鈴木由理　上田一人　瀧口伸雄　小谷善行　西村恕彦
（東京農工大学　工学部　電子情報工学科）

1．はじめに

自然言語処理において未知語の存在は大きな問題である。すべての単語を辞書に登録するのは不可能であり、仮にできたとしても新しく作られる単語は必ず未知語になってしまう。未知語というものが無くならないものである以上、これに対する的確な対応が求められる。

本稿では未知語の品詞を語形および文字種で推論する方法について述べる。

2．対象とする未知語

本研究では、文の形態素解析が終了している場合を想定して推論を行う。また未知語の品詞は次の八つのどれかであるとして推論する。

動詞　形容詞　形容動詞　名詞
副詞　連体詞　接続詞　感動詞

これ以外の品詞の単語は既知としている。用言はその活用形についての推論も行う。

3．推論方法

本研究では、二種類の推論方法を組み合わせて用いる。以下にそれぞれの推論方法を説明する。

3．1　語尾による推論

日本語の場合、動詞，形容詞，形容動詞（いわゆる用言）の三つの品詞は単語が語幹と語尾で構成されている。語幹は単語ごとに不変だが、語尾は単語の活用行，活用の種類，活用形に依存して変化する。未知語に語尾がある場合、その語尾がどんな文字列であるかによってそれらを推論できる。

Inference of Parts of Speech of Unknown Words from Word form and Types of Charactors

Yuri SUZUKI, Kazuhito UEDA, Nobuo TAKIGUCHI, Yoshiyuki KOTANI, Hirohiko NISIMURA
Tokyo University of Agriculture and Technology

3．2　接続による推論

付属語である助詞，助動詞は接続できる品詞（活用語の場合は活用形も）が文法上限定されている。もし未知語の後ろにこの助詞，助動詞があれば、未知語の品詞はその助詞，助動詞が接続できる品詞の中のどれかということになる。この推論は助詞，助動詞の品詞接続表を用いて行う。

3．3　二つの推論結果の組合わせ

仮に語尾による推論結果の品詞候補をG、接続による推論結果の品詞候補をSとする。語尾による推論と接続による推論の結果は絶対的なものなので、これらの推論結果から外れた品詞は、未知語の品詞としてはあり得ないものである。そこでGとSの共通部分をとり、品詞候補の幅をさらに狭めた。

例えば「行く」＋「けれど」という句は次のようになる。

G＝｛カ行五段動詞の終止形・連体形，
　　形容詞の連用形，用言以外の自立語｝
S＝｛動詞，形容詞，形容動詞の終止形｝

よって、G∩S＝｛カ行五段動詞の終止形｝

4．推論結果の重み付け

前述の推論方法を用いると未知語の品詞候補をかなり絞ることができるが、必ずしも一つに特定できるとは限らない。そこで推論結果として得られた品詞候補の中で、品詞xの信頼性を次式によりパラメータで重み付けする。

$P_0(x)=100/N$　…(a)
$P_n(x)=P_{n-1}(x)+W/N$　…(b)
$P_n(x)=P_{n-1}(x)+100/N$　…(c)

$P_n(x)$：品詞xのパラメータ値
N　：|G∩S|の数
W　：間違っている品詞が持っているパラメータ値の合計

式(a)は初めて現れた未知語に適用し、後の二つは二回以上出現した未知語に適用する。式(b)は、同じ未知語でも後ろに付いている助詞，助動詞が違う場合に用いる。式(c)は一度現れた未知語が別の活用形で使われている場合に用いる。このような場合には一回目の推論で狭めた品詞候補の幅をさらに狭めることができる。

5．文字種並びのパターンによる重み付け

式(a)では品詞候補の数で１００を等分しているが、文字種並びのパターンを調べることによって別の重み付けも行った。

日本語は表したい内容によって、文字の種類を使い分けている。その使い分け方には、ある程度規則性がある。例えば外来語（特に名詞）はたいてい片仮名書かれる。また動詞は、「行く」や「走り出す」のように漢字と平仮名を組み合わせて書くことが多い。このような規則を一般的な形で求めることにより、推論結果の信頼性を求められると考えた。

我々はＵＮＩＸの仮名漢字変換ツールＷｎｎの基本辞書を元に、文字種並びのパターンごとの品詞の数を調べた。そしてそれを次式により、前述の二つの推論結果の信頼性を求めるのに使用した。

$$P_0(x)=100 * m(x)/M \qquad \cdots (a')$$

m(x)：文字種パターンが同じで品詞がｘである既知語数

Ｍ　：パターンが同じで品詞が候補中にある既知語の総数

次の例を式(a')および式(c)を用いて信頼性を計算すると表１のようになる。

例）①「すくう」＋「と」，②「すくわ」＋「れ」

①Ｇ∩Ｓ＝｛ワ行五段動詞の終止形，名詞，副詞｝

②Ｇ∩Ｓ＝｛ワ行五段動詞の未然形｝

表１．例）のパラメータ値

品　詞	①	②	計
ワ行五段動詞	3	100	**103**
名　詞	40	-	40
副　詞	55	-	55

6．実験

我々は以上の方法で未知語の品詞を推論するシステムを実現した。そして文字種並びによる重み付けをする場合としない場合の二通りについて、実際に推論を行ってみた。推論したのは５６種類の未知語と助詞，助動詞の組み合わせ８６組である。この未知語は[3]より無作為に抽出したものである。結果は表２のようになった。候補中に正しい品詞がない場合はなかった。

表２．実験結果（単位%）

	◎	○	△	×
文字種なし	46.4	25.4	21.4	7.1
文字種あり	46.4	32.1	0.0	21.4

正答　－◎：品詞を一つに特定できたもの

正答　－○：品詞に複数候補があり信頼性が候補中最大

準正答－△：信頼性が候補中最大（他にも最大のものあり）

誤答　－×：品詞に複数候補があり信頼性が候補中二位以下

7．考察

約８割の単語について正しい推論を行うことができた。また文字種並びパターンによる重み付けは、平仮名のみの単語では約２割が正しい方向に推論され、片仮名の単語はほとんどが正解であった。

8．まとめ

本研究により語尾による推論と接続による推論を用いて未知語の品詞を推論することに成功した。今後は形態素解析と未知語の処理の間で双方向の推論を行うことを考えている。

参考文献

[1] Shigeyuki USAMI, Noboru OHNISHI, Noboru SUGIE: A Robust English Sentece Parser which can cope with Unknown Words , International Symposium on Natural Language Understanding and AI , pp.58-65 , 1992. [2]西尾実　他編，岩波国語辞典，岩波書店，1981. [3]辰濃和男，辰濃和男の天声人語・自然編，朝日文庫，1993.